Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2015-030 October 1, 2015

Big Data Privacy Scenarios

Elizabeth Bruce, Karen Sollins, Mona Vernon, and Danny Weitzner

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Big Data Privacy Scenarios

Big Data Privacy Working Group

September 2015

Big Data Privacy Working Group Chairs:

Elizabeth Bruce (MIT)

Karen Sollins (MIT)

Mona Vernon (Thomson Reuters)

Danny Weitzner (MIT)

bigdata@CSAIL

MIT BIG DATA INITIATIVE

Acknowledgements

We gratefully acknowledge the many contributors to this Scenario Working Document. This includes all of the Big Data Privacy Working Group leaders, team members, and guides for their thoughtful efforts. A special thank you to Dazza Greenwood of MIT Media Lab and Simon Thompson from BT for creating the original template for the scenario summaries.

Big Data Privacy Scenario Contributors/Teams: Micah Altman (MIT), Elizabeth Bruce (MIT), David Dietrich (EMC), John Ellenberger (SAP), Dazza Greenwood (MIT), Maritza Johnson (Facebook), Lalana Kagal (MIT), Jake Kendall (Gates Foundation), Cameron Kerry (MIT), Ilaria Liccardi (MIT), Yves-Alexandre de Montjoye (MIT), Una-May O'Reilly (MIT), Michael Power (Osgoode Hall Law School), Arnie Rosenthal (Mitre), Karen Sollins (MIT), Simon Thompson (BT), Mona Vernon (Thomson Reuters), Evelyne Viegas (Microsoft), and James Williams (Google/University of Toronto)

Big Data Privacy Working Group Editor: Barbara Mack (Pingry Hill Enterprises, Inc.)

Table of Contents

Executive Summary

Karen Sollins (MIT)

The MIT Big Data Privacy Working Group launched a series of workshops beginning in 2013 to explore the challenges and possible technological solutions to elements of those challenges. As a successor to those workshops, the Working Group began to focus on a collection of real world scenarios and use cases, to illuminate the challenges more concretely.

The deeper question explored by this exercise is *what is distinctive about privacy in the context of Big Data.* Although privacy as a general issue in computing and communications remains a topic of significant attention and disagreement, in this effort we narrow our attention to the "Big Data" context, to understand more clearly the particular challenges and possible approaches that derive from the collection, pooling, and combination of vast amounts of data, specifically about people. This focus on people as the subjects of attention in the Big Data context is central to the definition of privacy, which itself focuses on control data, information and inferences about people and how that can or should be used, exposed, or otherwise made available.

We summarize here an initial list of issues for privacy that derive specifically from the nature of Big Data. These derive from observations across the real world scenarios and use cases explored in this project as well as wider reading and discussions.

- Scale: The sheer size of the datasets leads to challenges in creating, managing and applying privacy policies.
- Diversity: The increased likelihood of more and more diverse participants in Big Data collection, management, and use, leads to differing agendas and objectives. By nature, this is likely to lead to contradictory agendas and objectives.
- Integration: With increased data management technologies (e.g. cloud services, data lakes, and so forth), integration across datasets, with new and often surprising opportunities for cross-product inferences, will also come new "information" about individuals and their behaviors.
- Impact on secondary participants: Because many pieces of information are reflective of not only the targeted subject, but secondary, often unattended, participants, the inferences and resulting information will increasingly be reflective of other people, not originally considered as the subject of privacy concerns and approaches.
- Need for emergent policies for emergent information: As inferences over merged data sets occur, emergent information or understanding will occur. Although each unique data set may have existing privacy policies and enforcement mechanisms, it is not clear that it is possible to develop the requisite and appropriate emerged privacy policies and appropriate enforcement of them automatically.

The primary content of this report is a number of real world scenarios, resulting from discussion and then subgroup efforts within the Privacy Working Group. Each case was analyzed along a collection of axes: key stakeholders, data lifecycle, key systems, potential privacy risks, and existing best practices within the context of that scenario. The template was laid out initially by Dazza Greenwood of the MIT Media Lab and Simon Thompson of BT and can be found in Appendix A.

As a result of collating these scenarios, two kinds of points emerged across them. The first is a small set of common questions. The second is a list of categories of stakeholders. We summarize those here.

The key questions that arose are:

- What new/unique challenges emerge when it comes to managing privacy in the context of big data?
- How do we assess benefit vs. risk?
- How do we evaluate "harm"? Given that harm is subjective, difficult to quantify, and falls on a spectrum from inappropriate online advertisements to discrimination in setting insurance rates to life or death medical intervention, is it possible to evaluate harm uniformly and if so, how would one do that?
- How can we establish and assess trust among the stakeholders? What mechanisms/models do we have for understanding trust?

A table of the categories of stakeholders derived from the scenarios can be found in Appendix B. In addition, Appendix C demonstrates an application of these stakeholder categories to the first use scenario on MOOCs and OLEs.

The initial list of categories of stakeholders includes:

- Data subject(s)
- Decision-maker
- Data collector
- Data curator
- Data analyst
- Data platform provider
- Policy enforcer
- Auditor

Both of these sets of points are discussed in more detail in the companion technology-mapping document, and are provided here to identify crosscutting observations from the various scenarios. Although the currently identified set of potential stakeholders is listed here, it is important to recognize that privacy is a much more complex problem that concerns more than the stakeholders alone.

The Working Group explored seven use cases. This report presents three in their complete forms in Sections 2 - 4; those three cases are described briefly in the Executive Summary. In addition, in the final section of the report, in Section 5, summaries of the additional four cases are presented, because these were studied in less detail.

Use Case: Massive Open Online Courses (MOOCs) and Online Learning Environments (OLEs)

Any online learning situation provides an opportunity to record all the activities of everyone involved in the teaching experience, primarily but not exclusively students and teaching staff. MOOCs as a subset of online learning take this to new scales and often to new levels of automation as well as expanding roles in the collection of, responsibility for, and use of the data that derives from those teaching experiences.

In focusing on privacy in this context, one is concentrating on questions of which behaviors and information about individuals may be exposed in ways that they may find contradicts their models of privacy. The challenges arise at least in part from the new opportunities that MOOCs provide to collect, merge, and reason over educational data at a scale and with an ease not previously possible. The data may now be used in novel ways and involve new stakeholders including data curators, data platform providers, researchers, and those interested in novel approaches to pedagogy. The challenge is to achieve that in ways that respect the privacy of the individual student, perhaps the teaching staff, and possibly secondary people as well, such as parents and guardians, especially in the face of asymmetric power relationships. One aspect of the challenge is to understand the implications of privacy "violations" in this context. They may arise not only from the direct exposure of information about the individual that was neither intended nor desired, but also from more subtle concerns over discrimination, harassment, inaccessibility, or violation of other civil and human rights.

The contribution identifies a number of key insights into privacy challenges that arise in the MOOC and OLE arenas, including:

- The nature of the information being collected, including click streams, contributions to online discussions, forums, and questionnaires, as well as behaviors with respect to both accessing and submitting content (reading, watching online lectures or videos, attempts at doing homework, etc.);
- Tools and norms for expression of privacy policies, including current, future, aggregation, and integration with other data;
- The tussles in objectives among students, teaching staff, owners of the educational content, crowd or student provision of contributions (through grading or social networking facilities) to the experience of other students, institutional hosts, educational systems (such as municipal school systems or state university systems), researchers and analysts, and service providers such as data curators, data storage and analysis services;
- The nature of the potential privacy violation harms to the various stakeholders;
- Translation of the Family Educational Rights and Privacy Act (FERPA) into this increasingly rich, complex, growing, and evolving domain in which collections of educational data is collected, curated, collated and perhaps integrated;
- The fact that this is previous uncharted territory with social, legal, and moral challenges as yet not clearly identified, which is also evolving due to increased technological capabilities, often independently of privacy objectives and interests.

Use Case: Research Infrastructure for Social Media

The behaviors of individuals and groups online can provide the basis for significant deeper understanding and prediction of human behaviors and interests. The kinds of data that can be useful in gaining that increased "social" understanding range from the various contributions made by individuals such as text, photos, various kinds of streaming media and other information relating to the participants as well as logged information such as click streams, frequency and other patterns of access, etc. At present the majority of access to such social media information is primarily restricted to in-house analysis by social media organizations.

The question explored by this group is whether and how one might provide a "privacy" framework for such information, giving the subjects opt-in control of which information

about themselves can be made available for broader studies and wider availability of the information. The intention is that permission for use remains with the subjects, but by giving them the opportunities to share, richer, and larger studies can occur, with all the potential societal benefits that those studies might entail. The subject must be given control over both the granularity and types of the data, including both static data such as birthdate, address, job history and so forth, and dynamic data such as ongoing posts in various media. In terms of the stakeholders, there are three key participants, 1) the subjects themselves, 2) the social media organizations who will play the role of data collectors, often data curators, and data platform providers, 3) the data analysts, who may also play the role of data curators, if they provided added understanding (curation) over the data sets. There are two general approaches to making the data available. The first is to generate slices, on some regular basis, of the data that is to be exposed and deliver that to the analysts. The alternative is to retain all data on a controlled service with a clearly defined API, providing only constrained access to the data. The first gives the analyst more freedom to explore, but reduces the subject's ability to retain control, especially with respect to withdrawing from a study retroactively.

There are at least four contexts in which such a system must operate: legal, social, business, and technical. The challenge is that privacy must be respected in the context of all of these domains simultaneously.

The study group identified a list of risks or challenges to privacy that must be considered in such a scenario including:

- Unexpected inference resulting from the analysis;
- Unexpected harm due to modifications of the data platform, due to inferences, or to the nature of the research itself;
- Unpredictable bias in the resulting research based on bias in the self-selecting nature of participation;
- Unexpected correlation between the study subject population and the general population;
- Removal from studies after agreeing to participate;
- Control of downstream use of the data, beyond the original analyst agreement. This raises questions of provenance (who has touched the data and how might they have modified it), to how to enforce policies beyond the bounds of pairwise agreements, to identification and recourse for misuse, for starters;
- Responsibility for data breaches both by the social media provider acting as repository and curator and by the researchers and analysts;
- Finding the balance between privacy and publication of results;
- Management of informed consents;
- Automation of as much of this as possible, while understanding the risks that may be introduced through such automation.

The study also identified some key technologies that exist and some places where technologies are needed, but not yet available.

The scenario is based on a current collaborative study involving the Technical University of Denmark and the MIT Human Dynamics Laboratory.

Use Case: Data for Good: Public Good and Public Policy Research Using Sensor Data/Mobile Devices

The challenge faced in this scenario is to take advantage of mobile phone data (mobility data) with one of two possible objectives. The first is to model and predict outbreaks of epidemics and the second is to enable micro-targeting of individuals or groups of people with interventions in order to reduce or prevent outbreaks of epidemics. The geographic region of focus in this work is Africa. Of particular interest are people moving across areas where an epidemic may be more prevalent and those where it may be less so.

In addition to the two kinds of objectives, the study examines two distinct system designs or implementations. In all cases, the original data is collected by the mobile network operators (MNO). In one implementation, each MNO anonymizes and coarsens the data both spatially and temporally. Thus, for example, the time may be reported in 12-hour blocks representing day and night and location may be represented as particular regions where malaria is prevalent or not. The individuality of each record is retained. This enables the targeting of individuals through one of two means. The anonymized identifier is presented to the MNO, which in turn either provides access information to the analyst or acts as an intermediary conveying information between the analyst and subject. In the other implementation design, data is merged on a regional basis before being aggregated, so for example, the MNO might report that a specific percentage of the residents of one area spent a different specific percentage of nights in a different target area. This second design significantly increases the subject's privacy and reduces the possibility of re-identification or exposure, as well as reducing the accuracy and potential utility of the data.

This study identified a number of challenges:

- The scenario exposes a direct tradeoff between health risks (and possible mitigation) for the individual and personal privacy;
- The scenario also exposes a direct tradeoff between analysis capabilities and personal privacy;
- MNOs are generally not in the business of anonymizing, curating and providing data to other entities. In these cases, the analyst role is often taken on by national health ministries;
- The legal bases for privacy in Africa are complex and generally based in historical tradition from the countries that colonized them in previous centuries. Those Western and Northern Africa mostly derive from the French civil code, with explicit privacy frameworks and are closely related to the European Privacy Directive. Those such as South Africa that derive from the English common law tradition have much less concrete policies with respect to privacy. To add to this, as populations move from one country to another, they may also be moving from one privacy policy model to another;

The intention of this use-case study was to allow the group to elicit commonalities and distinctions among the cases that might allow us to generalize. That in turn also has provided the basis for a companion paper, which concentrates on current and near-term future tools to improve the possibility of providing privacy, while continuing to allow for Big Data analysis and the benefits that accrue from that.

Other Use Cases

The report concludes with a brief summary of the additional four use cases examined by the Working Group. These include privacy under conditions of integrating over diverse datasets, the creation and management of user consent over exposure and use of personal data, consumer privacy and retail marketing, and genomics and health.

Conclusions

From these scenarios we draw three categories of conclusions. The first is a set of common overarching challenges. In order of increasing complexity these are:

- Scale: The sheer size of both the data itself and the accompanying meta-data that is necessary to manage it and provide privacy policies is increasing.
- Diversity: With growth, we also see an increase in the types of data, interests of analysts or users of the data, and richness of privacy policies in these new scenarios.
- Integration: There is increasing pressure and opportunity to merge or cross-fertilize among these diverse datasets. This leads to results that may have previously been inaccessible, but that are exposed through perhaps differing integrated observations of the individual.
- Secondary subjects: Although much data is based on primary subjects, it may also, perhaps inadvertently also reflect on secondary subjects. Handling privacy policies for this more integrated situation is significantly more complex than the policies applicable to a single subject.
- Emergent privacy policies: With both the integration of datasets and the increasing capture of data about secondary subjects, there is also a need for privacy policies to reflect this emergent data. The challenge of how these new policies come into existence will play an increasingly important role.

These scenarios have provide us with a basis for an initial observation about the differing stakeholders involved in the handling of Big Data and the privacy policies applicable to them. We begin with the subjects themselves, perhaps both primary and secondary, and the decision-makers who set out to have the data collect and made available. We then identify a set of different stakeholders having to do with the collection, management and provision of the data. This includes the actual data collector, the data curator, and the data platform provider. We then identify three kinds of stakeholders involved in the activities of usage of the data, the data analyst, the privacy policy enforcer, and the data access auditor.

With these challenges and observations in mind, we also recognize that there are a number of open questions. These questions revolve around several key elements. The first is whether or not Big Data brings new challenges to the provision of privacy or whether it exposes existing problems perhaps more clearly. More importantly, are questions of risk vs. benefits tradeoffs. One of the challenges one faces here is privacy and the risk of violation of privacy is not binary and perhaps not even measurable. Thus, one is then led to ask about the harms that may result from different levels of privacy policies and/or the violations of those privacies. Finally, we are left with a set of questions related to trust, how it comes into existence, how it may evolve, how humans' trust can be modeled, and how trust may be supported technically.

We note that this set of observations, challenges and questions are only representative of what one might draw even from this limited set of scenarios. A broader study might lead to yet more challenges and questions.

1 Introduction

Karen Sollins (MIT)

The vast amounts of diverse data that are now being called Big Data present society with an extremely interesting set of challenges, ranging from how to use any one such data set for a wide and increasing set of opportunities. These may range from improved product recommendations to improved modeling of human mobility in regions of infectious diseases to many other points in between. But Big Data presents additional opportunities that include a broader and deeper understanding across such data sets. If one can merge mobility data with medical histories, for example, one might provide a much more accurate model of potential epidemics, depending on both mobility and prior epidemics of diseases to which immunities are developed.

At the same time, societies and communities are becoming increasingly concerned over the questions of who knows what about them and whether or not they have control over those data collectors and analyzers knowing things about them. The concern is captured in the word "privacy". The "problem of privacy" is in fact a complex and subtle one, with many challenges and often too few solutions to those challenges. One must ask questions such as, "Who is the subject of the data?" There may be a primary subject, but data about interactions may have multiple primary subjects. There may be secondary subjects, such as the parents or legal guardians of a child who happens to be the subject of the data. In addition, one can ask questions about who else is involved with the data in various ways, such as collecting or storing it, protecting it, "curating" it for accuracy and completeness, analyzing it, and so forth. One can also ask what policies should be applied to the data for controlling access to it, to meet any privacy constraints from a legitimate policy source. Or, how might that policy be enforced? Or how can one be confident (trust) that the policy is either being defined by a legitimate policy source or being enforced by a trust-worthy enforcer? And so forth. The questions of what is meant by privacy, who can define appropriate privacy and how that might be implemented are only now beginning to be examined, with significant progress in some areas and less advancement in others.

The challenge we face in the Big Data arena is at the intersection of these two driving forces, Big Data itself and all that it has the potential to provide, and privacy, as it becomes increasingly well-understood to be a design-driver for systems in the cyber-age.

The MIT Big Data Privacy Working Group concentrates on this problem domain. To that end, several workshops were organized by and held at MIT.[1] In addition, the Working Group took on two initial agenda items: 1) documentation of a set of scenarios in order to better illuminate some of the central challenges to providing privacy in a "Big Data" world;

---

[1] See workshop reports:

1. *Big Data Privacy: Exploring the Future Role of Technology in Protecting Privacy*, June 19, 2013. Available at: report. (http://bigdata.csail.mit.edu/sites/bigdata/files/u9/MITBigDataPrivacy_WKSHP_2013_finalvWEB.pdf)

2. *MIT White House Big Data Privacy Workshop: Advancing the State of the Art in Technology and Practice*, March 3, 2014. Available at: report. (http://web.mit.edu/bigdata-priv/images/MITBigDataPrivacyWorkshop2014_final05142014.pdf)

2) road mapping of current and near-term future technologies that have promise of addressing parts of the privacy in Big Data challenge. This document is the first of these.

Below in the remainder of this section we will summarize a number of conclusions we draw from the scenarios. These take three forms. The first is a set of issues that derive from the larger challenge. The second is a set of categories of stakeholders we extract from the scenarios. Finally, we conclude the introduction with a set of questions, which remain unanswered, but appear to be central to the problem domain.

1.1 Overarching Observations

In examining the use scenarios here, we can identify an initial set of significant issues on the consideration of privacy, which derive specifically from the nature of Big Data. These are also informed by wider reading and discussions on the topic:

- *Scale*: The sheer size of the datasets lead to challenges in creating, managing and applying privacy policies. Because the datasets themselves are of such increasing size, the management of the meta-data that reflects privacy policies about it will incur parallel growth. One of the challenges is that as datasets grow, efficiency will play an increasing role. That will also be true of the privacy policy management associated with the growing data.
- *Diversity*: As datasets become "big data," it will be increasingly likely that more and more diverse stakeholders will be involved. Each may come to the effort with his or her own agenda. With an increasing number of stakeholders with different responsibilities will also come an increased probability that their interests, agendas and objectives will less aligned with each other and hence their approaches to privacy policies will also be more divergent and possibly conflicting. Thus, privacy policy conflict resolution will play an increasingly important role.
- *Integration*: With increased data management technologies (e.g. cloud services, data lakes, and so forth), integration across datasets, with new and often surprising opportunities for cross-product inferences, will also come new "information" about individuals and their behaviors. The challenge is that reasoning, inference and other analysis tools will allow for the recognition or discovery of hitherto hidden facts (data) about the subjects. This raises a question of how to create and enforce privacy policies on this new "data".
- *Impact on secondary participants*: Much data about individual subjects tends to reflect on other people as well. This may range from people who "liked" a post to people who are mentioned in email or posts, to true secondary participants, such as family members or co-workers. One question that will become increasingly important is how to observe the privacy rights of these other people, who are not the primary subject of the data and may not be available to apply a privacy policy when that is possible. Even if these secondary people are available, it is not clear how to handle conflicting privacy policies in this domain.
- *Need for emergent policies for emergent information*: As inferences over merged data sets occur, emergent information or understanding will occur; this will be based as mentioned above on both simply merging data sets, but perhaps more importantly allowing for the exposure of previously hidden data that is only exposed in the merging of datasets. Although each unique data set may have existing privacy policies and enforcement mechanisms, it is not clear that it is possible to automatically develop the requisite and appropriate *emerged* privacy policies and appropriate enforcement of them.

1.2 Stakeholders

As the reader will see in the scenarios themselves, there are a number of key stakeholder categories that appear repeatedly. Not all cases will include all of these stakeholders. In some cases, individuals may play more than one stakeholder role. Thus, for example, the data collector and the data curator may be the same, or the data platform provider, the policy enforcer and the auditor might be the same. But other combinations are likely to be found as well. It is also important to remember that the privacy policies for a dataset may be defined by people in different roles in different situations and, in some cases, the policies may be defined by outsiders on behalf of one or more of these stakeholders, as for example may be true under a regulatory regime. Thus, it may be that on behalf of the data subject, the government requires certain privacy policies.

- Data subject(s)
- Decision-maker
- Data collector
- Data curator
- Data analyst
- Data platform provider
- Policy enforcer
- Auditor

This list was drawn from the scenarios and should only be considered representative rather than complete. Appendix B includes a table with definitions of each of these stakeholder roles. It is also considered at greater length in the companion paper on technologies. Appendix C demonstrates an application of these definitions to the first scenario on MOOCs.

1.3 Open Questions and Issues

In studying these scenarios, we are left with a number of challenging questions and issues:

- **Novelty:** What new/unique challenges emerge when it comes to managing privacy in the context of big data?
- **Tradeoff:** How do we assess benefit vs. risk? Part of the challenge in these domains is that that the risks and tradeoffs need to be evaluated, to the extent that they can be evaluated by metrics, both by different metrics and at different time scales. As an extremely simple example, the benefits of MOOC analysis may be to future students, while the risks may be to the subjects of the data, the students about whom data has been collected. A key stroke logging system may help current students if the teaching staff can get immediate feedback on how long it takes each individual student to complete a particular exercise, but it may be that systematic changes may only occur on a longer term basis than the period during which a particular student is involved with a particular course. At the same time, to the extent that the data can profile individual students in numerous ways both in real time and perhaps over the longer life-time of the data set, and perhaps in conjunction with the data from other courses the student has taken, their risks of violation of privacy may continue to grow, and definitely are unrelated to the benefits for future students. One of the challenges in this domain of metrics is that privacy is not binary. In part because it is contextual and in part simply because the privacy of some information is more critical than other information, this question of the

tradeoff or balance between benefit and risk is both complex at any instant and is a moving target.

- **Harm:** How do we evaluate "harm"? As mentioned above, the risk to privacy is neither binary nor necessarily stable. The deeper challenge is to understand the potential harm that may accrue from potential risks. In fact, we may need to turn this issue around. The question we may need to ask is, "Which harms are important to the individuals and in what contexts?" Thus, harms could be imagined on a spectrum from inappropriate online advertising to discrimination in setting insurance rates to something that is a life or death matter in terms of medical intervention. From that we might be able to consider whether there is some metric for evaluating harm generically, or whether any comparative evaluation can only be done in terms of specific harms. In term, from the identification of harms, we may also be able to identify the risks that would lead to those harms. This is another way of talking about the related topic from the security community: threats.
- **Trust:** How can we establish and assess trust among the stakeholders? What does it mean for the various stakeholders to trust or mistrust each other or sets of others? What models do we have for understanding trust? What are the current and predictable future mechanisms and technologies for establishing trust and how do they relate to the models in people's minds and perception? How is trust established and maintained? How does it evolve over time?

With all these questions and issues in mind, the remainder of this report presents the scenario analysis done by various subgroups of the Big Data Privacy Working Group from which we drew these observations, thoughts and questions.

1.4 Remainder of This Document

The remainder of the document focuses on descriptions of the scenarios as outlined by subgroups of the larger working group. The first focuses on MOOCs (Massive Open Online Courses) and OLEs (OnLine Educational systems). The second addresses the challenges in using social networking data for research. The third considers the use of mobile cellphone data to reflect human mobility into and out of regions of highly infectious diseases, especially in developing parts of the world. The final section of the paper summarizes a number of additional scenarios addressed by the group, but in less depth. They illuminate more of the breadth of the problem domain. The paper concludes with three appendices: A) the template developed by the group for organizing the individual scenarios, B) a more in-depth table of the stakeholder categories, C) an application of the stakeholder analysis to the first scenario about MOOCs and OLEs, as an example.

2 Privacy Issues for Data Collected from MOOCs and Online Learning Environments

Team: Una-May O'Reilly (MIT), David Dietrich (EMC), Lalana Kagal (MIT)

2.1 Abstract

MOOCs (Massive Open Online Courses) represent a specific type of Online Learning Environment (OLE), which can be deployed on Internet-served platforms that collect large volumes of granular behavioral information about students' learning activities. Some data reveal each individual student's detailed study behavior such as video usage, consultation of text or learning tools, and the sequence in which material was navigated. Other data include assessments, grades, and social interactions and communication on forums within the platform. Collectively the data can be linked to auxiliary demographic information such as age, sex, and socioeconomic status. It can also be linked, if not anonymized, to public online behavior. A general set of legitimate uses of this data includes education research, examination, and analyses that directly or indirectly help instructors teach and conduct student assessments. Some, but not all, of these use cases have commercializable models for parties beyond the platform provider.

2.1.1 Definition of a MOOC and the Scope of OLE and MOOC in this document

MOOC is an acronym (Massive Open Online Course) originating in 2012. The acronym has been short-lived, as MOOC has evolved into a noun with meanings falling outside the acronym. For example, today we see MOOCs that are not open to all comers and MOOCs that are only partially online, because they are integrated into blended learning or flipped classroom models.[2] MOOCs share history with ITS – Intelligent Tutoring Systems and other learning management systems, such as Blackboard and Moodles.

We are focusing on data and its related privacy and confidentiality issues in this document. No OLE platform collects exactly the same data, but wherever it is largely unimportant to differentiate each platform by its specific name, we will refer to them all as OLEs.

2.1.2 State of Data Privacy Organization

OLEs, and MOOCs in particular, at their current scale are relatively recent, so data privacy and access policies are emergent and dynamic. Policy makers range in governance scale from the federal government to platform providers, and further to institutional and independent content providers. De facto policies and interim policies that have been necessary to cover fast-paced OLE activity both exist. Furthermore, existing policies on data privacy have been interpreted in new circumstances. Policy committees and meetings

[2] Given this fluidity of the meaning of MOOC, some people reasonably dispute the origin of the widely recognized first MOOC, believing large scale online courses at the college level preceding Ng's or Thrun's at Stanford in 2012 to be valid examples. It is arguable that Coursera and MITX/edX examples are more precisely called "xMOOCs," while previous online learning courses, which are generally much more fluid in nature in terms of content deployment, are more precisely called "connectivist" or "cMoocs."

abound. Policymaking is at the information collecting, option drafting, and revision stages. There is a potential to leverage the experience from many other data domains and shape a strong national example. This will require input from data stakeholders, the legal community, and technology experts. The latter are important because they can advise on technical risks of privacy and confidentiality breaches, while also indicating the capabilities and potential power of new technologies.

2.2 Detailed Narrative

The Online Learning Environment (OLE) data privacy scenario is relatively straightforward compared to some other domains, such as health records or personal genotyping, for several reasons:

- Because OLEs are recent, there are few data legacy complexities.
- Because the number of platforms is modest right now, the kinds of data are enumerable and their formats are known. However, this will change.
- Because there are enumerable classes of stakeholders in the space and policy precedents in related domains, there is generally less divergence and/or disagreement on what a policy should cover and what the principles and should be.

2.2.1 Open Issues

- Recognizing the dynamic nature of control of the data and acknowledging that the circumstances around that control may change. The data is replicated and passed by the platform provider to the institution of the content provider. At this point, two parties have control. Hereafter, designated controllers may expand in number, or the control may be passed from party to party in stages. Different controllers have different interests in the data and allow various parties to access it under a diverse set of goals and agreements. There is no uniformity to institutional practices across the country. If a broader policy and set of practices were to be developed by government, their interpretation might still result in heterogeneous local practices.
- Defining and determining legitimate uses of the data and how these uses should be controlled in a clear, specific, and open-ended manner.
- Setting guidelines or stated policies related to the sale, trade, or sharing of this data in OLEs and MOOCs.
- Defining and determining the legitimate commercial use of the data, if any.
- Defining the role of technology in aiding the drafting and governance of policy.
- Anticipating commercial and educational activities around OLE data, as well as potential malicious activities, and considering what technology can do to support them (or prevent them), as necessary.

The tradeoffs for policy around data control and access include:

- Students' right to confidentiality, privacy, and access to their own data.
- Institutions' and content providers' right to access because of content provision.
- Platform providers' right to access because of service provision.
- The benefit of research, the research-motivated right to access, and the countervailing risk of identification.
- The potential linking of anonymized data with outside data.

- Commercialization opportunities that may be unforeseen or unanticipated by students who grant permission to collect and control their data.
- The reasonable limits of technology for privacy and confidentiality policy support.

2.2.2 Additional Privacy Concerns

- Forum discussions and data linkability.
- One common way to grade assignments is via peer grading in MOOCs, which may create power relationships and opportunities for misuse.
- Power dynamics may not respect basic rights, as they relate to the linked data or the textual information from the discussion forums. In addition, the MOOCs can present asymmetrical power dynamics. Consider the case of children and prisoners, where people within a system (educational, correctional) may be required to do things as part of the that system, or in this case, the MOOC, and they may be influenced to bend the rules, given the existing power dynamics. Therefore, this area needs additional protection, since MOOCs have the potential to enable coercion and power imbalance. There are free MOOCs and MOOCs focused on certifications and jobs. There is an asymmetric power relationship in some situations and when this exists, there should be separate regulations governing these MOOCs to ensure that the dynamics are fair and there is free will and clear consent.

2.3 Privacy Impact Assessment - The Specific Context of Scenario 1

2.3.1 Actors

Students: Users who take the course, complete the assignments, and receive a grade.

Teaching content providers: Faculty and teaching staff that provide the teaching material, monitor and support the discussions, and handle the grading.

Crowd Participants: At-large parties who might volunteer to grade or offer feedback on assessments, programming assignments, and so forth, but who are not students or core teaching staff.

Peer Graders: A specific case of students, in which students are expected to grade each others work in order to manage the grading at large scales, as occurs in some MOOCs contexts.

Institutional content provider: The institution behind the teaching content providers. Examples include an enterprise offering in-house learning platform, a university offering a MOOC, an enterprise offering product education for clients, or the general public.

Platform provider (e.g. Coursera, edX, Stanford U): A party that deploys the course on the Web via a platform. In some cases, the same party develops and maintains the platform. For example, edX is a not-for-profit organization that develops, maintains, and deploys a MOOC platform as a service with a consortium of university partners, including MIT and Harvard. Coursera is a commercial entity and has different university relationships. Open edX is an open source platform that any content provider can adopt and use for content deployment.

Analyst: A party who examines the data collected from OLEs. Analysts include researchers, their students (if the researchers are academics), and education technologists.

Teaching staff, platform providers, and institutional content providers may also act as analysts.

Data controllers: Data control of OLE data is not always centralized or stationary. Examples of data controllers include the platform provider and institutional content provider. Within each of these institutions, there could be multiple controllers. They may control the data at different times, or they may control it concurrently. For example, at MIT, the Office of Digital Learning receives the data, controls its distribution at one point, and then later passes this role on to the Institute Registrar.

2.3.2 Actors and Relationships

Analysts interact **with data controllers** to gain access the data. The data controller often asked the analysts to formally submit to a policy. Eventually analysts will transform source data by linking and interpretation into more abstract representations of student behavior, e.g. variables for modeling, all the while trying to enforce student anonymity. Analysts will interact with data controllers to work out how to more widely share such variables and to evaluate the risk that they and models using them present some risk of re-identification.

Data controllers interact with each other to pass or share the data.

Students interact with the **platform provider** and the **teaching content provider**. They register with the platform provider to gain entry to the platform and course. They provide background information, participate in the course, including its forums and assessments, and provide survey information. As data controllers, both providers will access this information. It should be noted that students often confuse the platform and content providers. A student is shown a privacy and access policy by the platform when he or she registers. A student agrees to a platform use policy when registering. For example, edX's use policy stipulates no scraping.

Students indirectly interact via the OLE with the **Institutional Content Provider** when they have grades placed in their academic records, or when they receive credit or proficiency certificates.

Students indirectly interact with **analysts**. They gain a benefit from assistance that could be founded on the researchers' analysis of their data – both a student's individual data and the data of other students in aggregate.

Students interact with other **students**, generating data of great interest. These interactions frequently take place on forums within the platform. Importantly, for data privacy reasons, they may take place *outside* the platform, informally arising, rather than being organized by the course structure. Examples of digital records of these interactions are Facebook or Linked In groups. Sometimes students assess the work of other students in peer-to-peer relationships. Students may also work in groups on projects or homework.

Students interact with **Crowd Participants** when they receive feedback from them. For example, one course at MIT invites alumni to comment on student software designs.

Students rarely interact with **data controllers** at this time and have zero or little access to their data beyond official records created for their education purposes.

Institutional content providers employ the teaching staff, i.e. **teaching content providers** and have agreements with them regarding intellectual property related to the course, and remuneration for instruction. The institution is usually the data controller,

rather than the teaching content provider. In fact, the latter party may need to seek permission for data access to the very courses he or she has taught.

Teaching content providers interact with **Crowd Participants** to provide guidelines on grading and get feedback on student performance and interest in the course.

Teaching content providers provide feedback to **institutional content providers** and **platform providers** on usability, additional features, and student performance, for example.

Teaching content providers may interact with **analysts** to understand how students learn and interact with their teaching content in order to improve that content.

Teaching content providers may interact with **data controllers** to get access to data about their course in order to analyze it and to improve the teaching content.

Institutional content providers interact with the **platform providers** to ensure that the courses are supported properly and provide feedback on additional features.

Institutional content providers interact with **data controllers** to identify and/or specify the policies that they wish to enforce and to discuss enforcement mechanisms.

The core interaction is the student learning via an OLE. Around this point, students interact with each other and teaching staff. In terms of privacy, students are identified by their login id on the OLE platform. They may also reveal their "offline" identify to each other and staff in the content of their discussion posts. Students agree to a platform use agreement that implies that they accept the platform's data use policy. During the learning process, the platform provider captures clickstream, assessment, discussion, and wiki data. In real time, or at longer intervals, the platform provider aggregates this data from many students' interactions. The platform and the institutional content providers control these data. They are generally not accessible to the student, but they are accessible to teaching content providers and analysts. Institutions controlling the data are responsible for meeting FERPA requirements and pseudo-anonymizing data to which they will link and provide access. They also develop and provide technical support for data access policies. Analysts transform source data in the course of their modeling activities. They may combine low level observations (e.g. mouse click activity) into variables (e. g. referrals to text during problem solving) and compile large datasets of them. These datasets describe student behavior at a recognizable level of human activity. They are destined to become the data "currency" of analytic research. How to handle the control and privacy protection of such secondary data (i.e. who can it be shared with, given potential for student re-identification) remains to be resolved.

2.4 Goals of OLEs

General: To educate. With college OLEs, the education could have (secondary) outreach, accessibility goals. With corporate OLEs, the education could have (secondary) product adoption, sentiment, and publicity goals. In addition, goals specific to actors are:

Teaching content providers: Providing teaching materials, job tasks for an employer.

Crowd Participants: Altruistic or professional education goals.

Peer Graders: Evaluate other student work in an appropriate, objective manner.

Institutional content provider: Sometimes through generating revenue directly or indirectly; reputation.

Platform provider: Revenue streams via advertising, signature tracks, recruiting. Possible cross selling to steer people toward formal degree programs at universities that provide content. Own the ecosystem, as they own the actual platform and access the data.

Analysts: Research into education, improvement of OLE experience for students and teachers by interpreting historical data. Inevitably, financial profit could be a goal for this kind of actor.

Data controllers: These are the data gatekeepers. They regulate access to the data at the moment for analysts and other potential controllers. Their goal is to ensure that the privacy and confidentiality policies governing the data are respected, while providing access to appropriate analysts.

There is a lurking unnamed adversarial goal/actor in this space: Those exploiting the data for commercial or hacking purposes, outside the realm of educational use, i.e. to identify someone and target her or him specifically for revelations or for profit-based activities. For example, there is a significant potential for targeted advertising.

2.5 Data

MOOCs offer a potential social science laboratory or study setting where students' behavior and interaction with course content can be almost microscopically observed. Technology allows us to capture a tremendous amount of detailed data, including:

- Click-stream interactions between a student and content.
- Use of videos and other e-resources, such as digitized reference material, wikis, and forums.
- Assessment behavior: attempts, correctness, use of immediate feedback.
- Self-reported background, pre- and post-test surveys.
- More data than in a residential setting, but with less contextual information accompanying it.

This data can be segmented in several ways, as outlined below.

2.5.1 Course-related

- Course content from content provider.
- Data exhaust from platform, as students interact with Web servers. This is often called clickstream data. For edX, it is JSON logs of every get/post of data to the web site.
- Student input to the OLE via wiki and discussion forum entries, questionnaires, and self-reporting surveys.
- Assessments – both grades and responses; certificate achievement.

2.5.2 Institution or Platform-related

- Curricular data related to courses taken, timing, and learning paths.
- Registration data, such as profile information about students.
- Payment data perhaps (e.g., Coursera Signature Tracks, other third parties).
- Certificate data.

These data are in diverse formats and can be linked to form student-oriented or time-oriented descriptions (the former being more actionable) of learning activity *within the platform*. One such open organization of MOOC platform data is MoocDB within The MoocDB project. MoocDB is a platform agnostic functional data model for data exhaust from MOOCs. The MoocDB project will provide open source software of MOOC tools and frameworks.

2.6 Systems

Business systems. As an example, Coursera is a for-profit organization, providing an online service. In the past, Coursera offered a "Freemium" model in the marketplace, and has evolved to offer low cost courses and specializations. Signature tracking verifies student authenticity, recruiters are in the model and serve as a revenue source, and lifelong learners take courses well beyond the typical student years. In the case of a corporate MOOC, HR learning systems are part of this picture as well.

2.7 Risks

The biggest data risk is that someone in the data is identified and this causes harm to them. Data has to be pseudo-anonymized before release, but that does not assure that re-identification will not be possible with 100% confidence. Re-identification can take place in at least three ways:

- Pseudo-randomized data has confidential cross-reference tables to true identity. These tables, if not adequately protected, could be compromised.
- Some reference in the content of the data, for example free text posts in discussions or timestamps will directly or indirectly allow cross-referencing to public data that reveals identity.
- A previously compromised dataset can potentially be used to learn the behavior of a student and this behavior pattern can then be applied to new datasets to identify the student.

Several additional risks exist:

- Data control is not in the hands of the data providers, i.e. the student. Therefore, there is a risk that the data can be used in a way that the data provider did not anticipate, or for a reason that they do not approve.
- Data released for research purposes will be used for commercial purposes.
- Data will be used to evaluate the teaching ability of the teaching content provider and to compare teaching content across different institutional content providers without explicit consent related to individual data sharing.
- Students may not understand the privacy policy that they have agreed to at sign-up, and their personal data gets shared or monetized without their informed consent.

2.8 Rules/Regulations

In the United States much of the regulation of academic data is regulated by the Family Educational Rights and Privacy Act (1974),[3] which defines the rights of parents and

[3] See http://www2.ed.gov/policy/gen/guid/fpco/ferpa/students.html for general information about FERPA.

guardians to access and some control over who has access to which information about children under 18 years old. It also defines the rights of students over 18, such as students in college. It is important to recognize that there may be a number of non-FERPA regulations with respect to the privacy of information about students. An example of this is the U.S. Health Insurance Portability and Accountability Act (HIPAA), but there are others as well. This group did not discuss the relationships among these various different factors in the privacy of educational data, but just noted that such differences and possible conflicts exist.

2.9 Technologies

- Learning Platforms (using this broadly to refer to platforms such as edX, Coursera, Udacity, and other MOOC providers, as well as more traditional Learning Management Systems (LMS) such as Blackboard;
- Software frameworks for processing large datasets, such as Hadoop and data lakes that store a combination of structured and unstructured data;
- Web browsers and front end tools;
- Analytical tools;
- Cloud computing platforms (e.g., Amazon Web Services and others);
- Code on different systems;
- Mobile devices.

2.10 Privacy Constraints

Privacy constraints in a MOOC are very different from those of a physical classroom experience. There is a perspective that since MOOCs are much more open, students are more vulnerable online, compared with a traditional classroom setting.

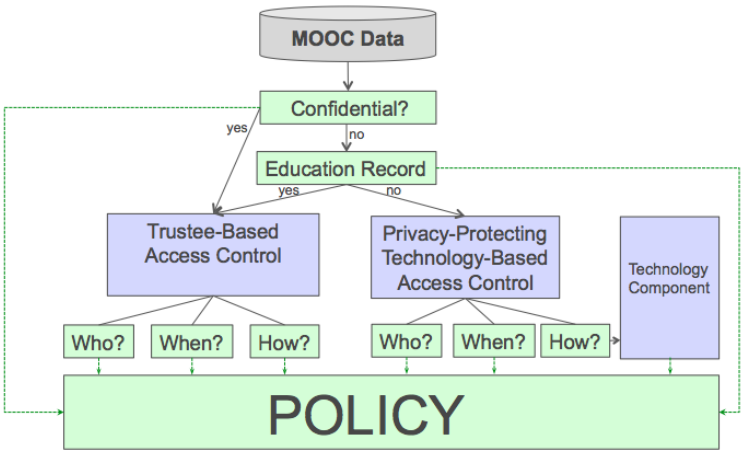
2.11 Technology Informing and Supporting OLE Data Privacy and Confidentiality Policy

2.11.1 What tools and approaches can (new) technology provide?

Some possible technologies:

- Differential privacy.
- Analysis is carried out on encrypted data, so even the platform provider does not see the data (homomorphic encryption).
- The analyst uses trusted and privacy-aware API to write up their analysis and submit their code to data controller; the API prevents the abuse of data.
- Store extensive audit logs about analyst access to ensure that the analyst is not able to chain queries in order to gain access to inappropriate data.
- Privacy-aware analysis framework that helps analyst be policy compliant.

Some initial thinking has been given to managing MOOC data via decision and policy engines based on heuristics. This approach would require separating the databases and using different access controls.



2.11.2 Risks

What risks are there to even the new technology?

- Differential privacy only works within a closed dataset; privacy breeches are possible when external datasets are linked.
- Encryption acts like access control and is useful when the platform provider is untrusted.
- A restricted API acts like an access control combined with audit.
- Auditing can handle post facto problems.
- The analyst platform provides a holistic approach to access control, privacy awareness, and ensuring policy compliance. However, it restricts the analyst to a single platform.

3 Research Infrastructure for Social Media

Team: Maritza Johnson (Facebook), Dazza Greenwood (MIT), Mona Vernon (Thomson Reuters)

3.1 Abstract

Most social media platforms provide at least two basic features: the ability to share user-generated content and the ability to connect with an audience. Different social media platforms make it possible for users to share a range of content types and some allow the user to selectively choose the audience for individual pieces of content. On Facebook, for example, the user could share text-based status updates, photos, or websites URLs. The user is also able to comment on content posted by other users, install applications that utilize the Facebook API, or communicate with others within a self-organized group of people. Between the user-generated content and the server logs that capture how and when people interact with the platform, these services are an invaluable source of information about human behavior at the individual, group, and even country levels.

The goal of this scenario is to evaluate technical solutions that would open this data up to researchers while offering data subjects informed consent and control over their data. Studies of social media to date have provided insights on topics as wide-ranging as social capital, social influence, meme evolution, emotional contagion, mobility, and politics. For a variety of reasons, much of this research is currently limited to employees of social media companies.

3.2 Scenario Introduction

Studies of social media to date have provided insights on topics as wide-ranging as social capital, social influence, meme evolution, emotional contagion, mobility, and politics. Unfortunately, much of this research is currently limited to in-house researchers at social media companies. Academics and other researchers have, in some cases, leveraged publicly available content or APIs, when they are available, but there are notable limitations to collecting data through these channels. In some cases, studying a group of people yields the most interesting insights but this requires that a critical mass of the population opts-in to a research program. In other cases, the user-generated content is best supplemented by information that can only be found in the server logs, such as how frequently a person visits the platform, how much time they spend, and the proportion of time spent consuming content versus producing content.

One way to increase the volume of research in this area is to develop a social media research infrastructure that allows users (data subjects) to opt-in to a program that makes some subset of their social media content and the accompanying server logs available to researchers. The result would be a large-scale, rich dataset that would empower researchers to generate varied and reproducible research. Social media platforms might participate in data release program with varying options. For example, one successful implementation of the program might include a predefined set of user data and data from server logs, a feature that allows researchers to contact participants for supplementary data or follow up surveys. It might also include a portal with educational content for individuals to visit to help them understand the information they've chosen to donate, to see how researchers are using it, and to gauge the long-term benefits of participation.

The incentive for the Study Participants and Social Media Providers is to act for the public good. The risk for the Study Participants is that they might experience negative effects as a result of contributing their data to the general dataset. The data exchanged may contain several features of data known to be personally identifying or sensitive in nature including race, sexual preferences, gender choice, and political views. The data exchanged could also be used for making unexpected inferences that the participant was unaware of at the time of consent.

As a high-level overview, the program would be initialized by the Social Media Provider. The social media provider would advertise the opt-in research program to users (potential participants), give an overview of the structure of the program, the risks, and the benefits and present the choices that represent how a user might participate. This information would include the main features of the program: the basic set of information that is required to participate; additional optional fields that the participant may choose to include; and the features that would allow a researcher to contact a user for additional information.

The participant will have granular opt-in choices for sharing a subset of their personal data, for example, some basic (**static**) fields are included in the set such as birth month and year, current city, school history, job history, etc. The participant is also given the ability to contribute **dynamic** streams of their data, including photos, posts, comments, and interests.

The information will clearly describe the policies that researchers will be held to, while making it clear that the dataset is not believed to be anonymous or de-identified in a robust manner.

3.3 Stakeholders and Interactions

Social media providers are the **data collectors** and would initially serve as the **data platform providers.**

Social media users are the **data subjects** and are asked to provide informed consent for the data to be transmitted by Social Media Provider to Researcher for purposes of research study.

Researchers are **data analysts** and receive data from data collectors (social media providers) by permission of the data subjects (social media users). The Researchers become **data curators** of the data that they receive at the time of receipt and any derivative data that is produced as a result of the research activities.

The data collectors (social media providers) remain **data curators** for the underlying data of all social media users that they continue to maintain.

Social media users will continue to interact with the social media platform to generate new content.

Researchers might contact social media users to collect additional data to supplement the social media data.

Social media users will contact the social media provider if they experience issues or have concerns about the overall program. Users will expect that the social media provider is ultimately responsible for ensuring a positive experience.

Researchers would provide information to the data subjects about the research that results from using the data subjects' data.

3.3.1 Data

Examples of the data that could be made available:

- Posts: photos, status updates, location check-ins, etc.
- Comments and the number of likes on individual posts
- Education history
- Hometown
- Current city
- Religious and political views
- Information about the friend network: summary statistics like count, breakdown by age range, current city (location), gender, political views, and education level, etc.

For the dynamic fields, the informed consent dialog might offer the ability to contribute:

Audience, keyword, tags, or some other mechanism could define the exceptions.

- All historical data
- All historical data with some exceptions
- Only future data
- Only future data with exceptions
- Historical and future data
- Historical and future data with exceptions

Making the data available:

Option 1. Social media provider generates data slices:

- On a monthly/quarterly/annual basis, the Social Media Provider would create a new data slice for all active participants in the program.
- Participants would be able to opt-out of the program, but they would not be able to remove their data from the datasets that had already been published. *This is mainly because no practical guarantees could be made about deletion requests once the data has been released to researchers.*
- Researchers would conduct queries on the available datasets, or download the entire available set for a given time period.

Option 2: Social media platform provides as API specifically for this program.

3.4 Systems

Legal systems – The privacy policy, or data use policy, currently governs how data can be used.

Social systems – What are the existing expectations around who owns shared content? Social media data some times involves more than one data subject. Consider for example a Facebook status update with a set of comments and "Likes." The simple text of the post belongs to the original poster (the person we consider the data subject throughout this scenario). But the post might also include "tags" to other people. These structured references to other users represent other individuals. What's the best way to handle

providing this information in the dataset? Similarly, on Facebook, comments on a post in are stored with the account of the post author rather than the commenter. Who does this content belong to? The comments are relevant to the context of the post, but are generated by other people. Is consent required to know which users "liked" a post? Do we limit the data so that only the number of likes is available?

Business systems – Human subjects research requires the approval of an ethics committee if The Common Rule[4] applies.

Technical systems – informed consent, a permission-based system to allow the user to participate in a way they feel comfortable, transparency and control over how data is shared, deletion protocols, de-identification of data to protect individuals when it is aggregated, and auditable systems to understand who has access.

3.5 Analyze the Scenario

3.5.1 Goals

- The participants benefit from contributing to a general body of knowledge and perhaps they will learn something about themselves on an individual basis too.
- Researchers have access to a dataset that was previously unavailable.
- The social media provider gains insights about the user base and contributes to the general body of knowledge.
- The Researchers may be acting for the public good, or they may be acting to develop their own careers.

3.5.2 Risks

- Participants agree to participate in the program and then later experience an unexpected harm, due to an unexpected inference that arises from the research.
- Participants agree to participate in the program and then later experience an unexpected harm, due modification of the site based on those inferences, or as a part of the experiment itself.
- The dataset would be a valuable resource for researchers, but it would be difficult to quantify the bias introduced to the dataset based on the characteristics of the people who decide to opt-in to the program.
- Researchers identify a correlation in the study population that can be extrapolated to the general population, greater than the pool of the participants who opted in.
- Deletion requests -- is it reasonable to design the program such that people can opt in or choose to opt out, but cannot remove their data from the already-released data slices? If not, then how would deletion be handled when the data slices have already been released?
- Lack of control on the downstream use of the data, or derived data: what are expectations and commitments to the people who opt in on downstream uses of the data? When new insights emerge, how do you ensure that the inferences/derivative data have been created in a way that is consistent with an individual's

[4] The Common Rule is the name of the U.S. federal policy on the ethics of use of human subjects in biomedical and behavioral research. For more detail see http://www.hhs.gov/ohrp/humansubjects/commonrule/index.html

expectations? How would we detect a misuse of the data? How would we tag derivative data to understand where it came from and understand the original policy in order to determine whether the action and the future uses are policy compliant?

- The data copy may be disposed of by the Researchers after the study, or may be retained in a corpus for further study. The data copy must be held securely and the Researchers are liable for a breach. However, the Social Media Provider may be liable if they have not assured that the Researchers are acting properly and also may risk collateral damage in the case of a breach, even if the proper processes have been followed. A variety of technologies and systems will be used to store and transmit the data, including Internet links and various databases. The data must be held according to the various Data Protection regulations in the territory that the data has been exported to, provided the export is legal in the first place.

3.5.3 Rules

- Terms and Conditions of the social media provider
- The social media platform's existing audience controls for content
- Notice and consent when the user opts-in to the program
- FTC Section 5
- For the Researchers: applicable human subjects research protections (e.g., The Belmont Report or The Common Rule)
- The policies of publication venues

3.5.4 Time

Roughly two to four years.

3.5.5 Existing Relevant Best Practices

Human subjects review committee -- Where The Common Rule applies an ethics committee would be required to give approval for human subjects research and an appropriate risk assessment would be undertaken to validate the arrangements that have been put in place to manage the data security and disclosure.

OAuth2 for enabling access to authorized users -- Once the data subject has provided the click-based grant of authorization, the Researcher could be granted an OAuth2 token to request and receive that individual's data via the API. The data would then be transferred to a research platform and database to conduct the analysis. The OAuth2 token would be provisioned to include authorized access to a scope of access that corresponds to the personal data that the data subject agreed to provide.

In the UK, organizations like the UK Data Archive can be consulted to manage the privacy processes and publication of results without breaching privacy.

3.5.6 Gaps

The above description includes a few caveats that are based on the limitations of our technical abilities -- for example, it's important that the participants understand that researchers would agree to a policy that prohibited attempts to re-identify participants within the dataset, but it would be difficult to make any guarantees along those lines given today's technical solutions. Similarly, there could be contractual limitations in place

around deletion and retention, however, we are lacking technical systems to enforce the policies.

- The management of access to data and the risks associated with publication present an impediment to the use of social media data.
- Gathering informed consent from social media users is particularly problematic.
- To enable research of this kind, we need to streamline these processes and provide automatic verification of the safety of disclosures.

3.6 Innovation Ideas and Opportunities

3.6.1 Looking at 3-5 years opportunities and challenges

One of the main opportunities lies in the ability to combine social data from different sources in order to conduct more insightful research and enabling reproducibility of research. This will require technology to allow for privacy preservation, or the application of rules as the data is combined with other datasets.

How do we develop legislation, if it is not already in place, to set-up a baseline that will not be country-specific and hence makes it difficult to manage for the social media providers to comply to multiple forms of legislation? Ideally, there will be a mechanism for allowing social science research to be conducted on a global scale.

The essence of computational social science may become more common and "normal," compared to the niche role that computation currently has in the social sciences. A true limitation of the research area now is that only social media platforms have easy access to large-scale datasets. Most academics who work in the space have partnerships with corporate entities to acquire large datasets. How will the research community change when large-scale datasets are available to all social computing researchers?

Shifting norms are expected to continue even beyond the 3-5 year horizon and this means that we expect continued deep uncertainty.

3.6.2 Open Questions

- What if we developed a "Common Program & Protocol" for infrastructure-level services to enable population-wide Living Labs social media research?
- What if Facebook supported a feature for users to "opt-in" for participation in pre-qualified research studies and we modeled/tested that as a common service available to any approved MIT Living Lab application? In theory, this sort of capability could enable re-usable or easy update of consent across simultaneous research studies and for future studies. This type of service could comprise fundamental capabilities that are now missing for operationalizing fair permission-based use of personal data in big data contexts.
- An OAuth2 scope type developed for research content could be a model for other social networks to use. One of the best aspects of the Facebook and general Web 2.0 design pattern with OAuth2 is that the authorizations can be seen on a dashboard and individually modified or revoked according to the agreements, potentially at any time.
- How could a common service type and interface specification be used by researchers to enable other social media providers (e.g. LinkedIn, Google Plus, Twitter) to provide consent-based data using interoperable programs and

according to the standard protocol developed by MIT and Facebook? What issues of scaling, cost/risk management, business value, and usability would need to be addressed, and at what phase of design, development, testing, iteration, and deployment (alpha, beta, V1, V2)?

- Could MIT Living Labs partner with Facebook to test a model Open PDS (Personal Data Store) deployment that further developed infrastructure-grade service interfaces, pipes, and gauges? Would - or should - it matter if Open PDS was situated at the research institution (e.g. MIT for MIT Living Labs), or at a third party provider?

3.6.3 Alternative A: Interactions of People

The participant has an account with Social Media Provider, provides Informed Consent to Participate in the Study and, within the scope of the study, provides authorization to Social Media Company to release personal data to Researchers via their applications.

A laboratory has an approved research study and has received the informed consent of individual participants and has registered an application with a Social Media provider and selected the OAuth2 scopes for grant of authorized access that correspond to the personal data used to conduct the research. Once the Individual has provided the click-based grant of authorization, the lab's App uses an OAuth2 token to request and receive that individual's personal data via their app and into a research platform and database used to conduct the analysis.

The Social Media Provider provides an account to the individual under its terms and conditions and provides a developer account to the lab under another set of terms and conditions. It also provides the personal data authorized by the individual for sharing with the application of the lab upon permission of the individual user.

3.6.4 Data

All past and current available data during the course of participation in the study that is available by OAuth2 individual consent from included social media providers.

3.7 Notes on Scenario

This example is based on a study that is currently happening at the Technical University of Denmark in collaboration with the MIT Human Dynamics Lab. However, references to potential downstream sharing arrangements by Participants and Researchers represent prospective future phase research and assume a future state of perhaps 1-3 years from now.

3.8 References

Related to applicable rules

** When Facebook has the data, these terms apply:*

Platform Policy (Applies via Researcher's Registered "Client" App/Service)
https://developers.facebook.com/policy

Statement of Rights and Responsibilities
https://www.facebook.com/legal/terms

Data Use Policy
https://www.facebook.com/about/privacy

Facebook Community Standards
https://www.facebook.com/communitystandards

Facebook Principles
https://www.facebook.com/principles.php

** When the Researchers Receive the Data*

SensibleDTU Example Computational Social Science Research Study
https://www.sensible.dtu.dk/?page_id=89

** When the Participants Share Downstream Via Personal Data Services*

MIT Human Dynamics Lab Model Personal Data System Rules
https://github.com/HumanDynamics/SystemRules/blob/master/Model_Personal_Data_System_Rules.md

Draft Data Rights Services Agreement
https://github.com/HumanDynamics/LegalAgreements/blob/master/DataRightsServicesAgreement.md

4 Data for Good: Public Good and Public Policy Research Using Sensor Data/Mobile Devices

Team: Jake Kendall (Gates Foundation), Yves-Alexandre de Montjoye (MIT), Cameron Kerry (MIT)

4.1 Abstract

There is little doubt that the capacity to collect and analyze mobile phone data at large scale has great potential for good [UN][D4D]. There are, however, numerous barriers that need to be overcome before this data can be broadly used by non-governmental organizations (NGOs) and researchers:

- The data is generated by the carriers' infrastructure and belong to them
- The infrastructure to manage and analyze this data at scale for good has to be developed
- Data-science skills are needed within NGOs to fully take advantage of the data,
- These data are highly sensitive and personal - simply anonymized mobile phone metadata has been showed to be re-identifiable [unique], and
- The legal and regulatory environment is at best uncertain and may prevent certain uses of the data.

This group is studying the technical and legal solutions that could make this data available in an operational context. We first focus our analysis on two scenarios inspired by the available academic literature. We then sketch proposed practical implementations to operationalize these scenarios and analyze them from a privacy angle, focusing on re-identification, and a legal perspective, with a focus on African countries.

4.2 Scenario Development

After considering a number of different scenarios, we focused on two that contrast scope and purpose:

Scenario 1: Tracking population mobility within and across borders to model epidemic spread

Scenario 2: Micro-targeting behavior change interventions to individuals or specific sub-sets of the population.

Scenario 1 is modeled on the use of location data coming from mobile phones in order to better understand and quantify the spread of malaria. The location of users is recorded at the antenna level and every time a user is interacting with his phone (phone call, text, or Internet session), location data is used to estimate his migrations between a set of predefined regions, for example from Nairobi to Lake Victoria, as well as the total number of nights spent by every user in every region. The main expected outcomes of this work are two matrices that show the average monthly parasite importation by returning residents and by visitors. In the scenario we consider, such matrices would be computed on a monthly basis and shared with local CDCs, ministries of health, and NGOs. We also consider a case where data from multiple operators across neighboring countries would be used to estimate the monthly parasite importations per regions. While this scenario has a clear public purpose, the sensitivity, re-identifiability, and potential for misuse of fine-grained

location data, such as targeting of individuals or groups for malicious purposes, has to be considered.

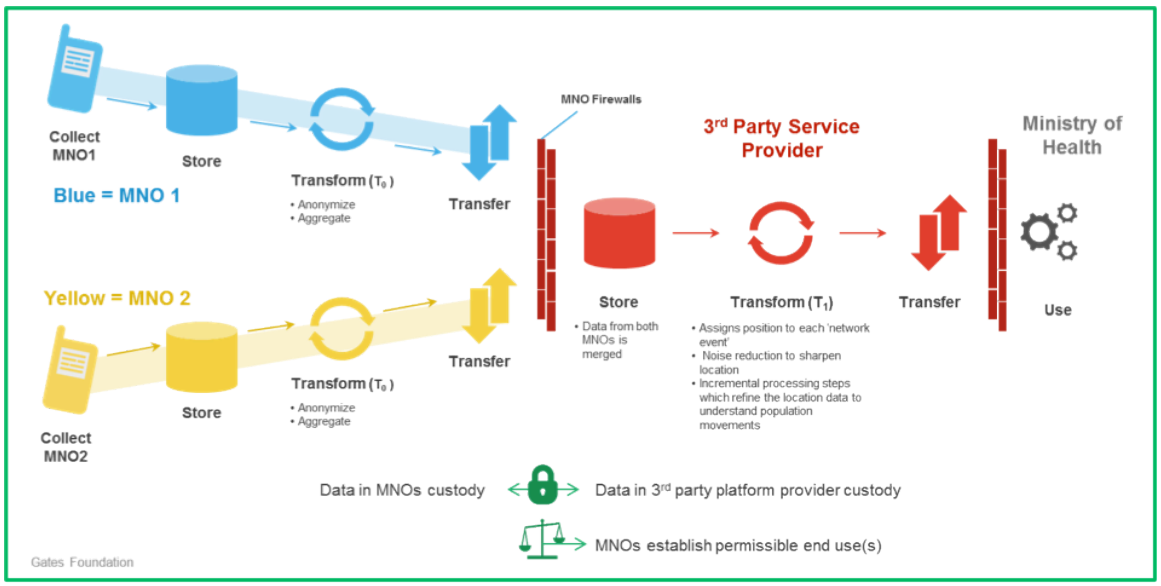
Scenario 2, inspired by [bigdatadriven], uses mobile phone metadata to micro-target people for specific behavior change purposes: agriculture techniques and health seeking behaviors, for example. In this case, location data at the antenna level, as well as other metadata fields, such as anonymized call and text logs (excluding content), and recharge information are used to estimate an individual's status (farmer, other socio-economic status) and/or propensity to change behavior. In this scenario, mobile phone metadata are used by machine-learning algorithms through a set of pre-computed metrics (e.g. daily distance traveled, recharging behavior, time it takes to answer a text,). Users can then be targeted for various behavior change or informational campaigns through text messages or phone calls sent by the carrier, or by a third party. While computing the metrics requires a rich set of data, this scenario aims at emphasizing the challenges associated with micro-targeting individuals and in introducing an element of intrusiveness that is not present in Scenario 1, but involves the same public purposes.

4.3 Operation of Scenarios

For each scenario, we propose two potential implementations. We will subsequently analyze these four implementations from a privacy angle and a legal perspective.

4.3.1 Scenario 1

In Scenario 1 implementation A, the different mobile network operators (MNOs) involved would share simply anonymized individual mobility data with one third-party. To limit the risks of re-identification, the data would be coarsened spatially and temporally. Matching the study [quantifying], the spatial resolution of the data would be at a predefined regional level or approximately 1000km^2 (692 settlements for the 581,309 km^2 of Kenya). Similarly, given the importance of nights for malaria infections (mosquito bites), the temporal resolution of the data would be of 12h (e.g. 6am-6pm). Finally, as malaria symptoms may take up to 30 days to manifest themselves, we work under the assumption that three months of such mobility data are needed to estimate the impact of human mobility on malaria. Different MNOs would hash a slated version of the mobile phone number of the subscribers to allow the third party to reconcile the data. Scenario 1 implementation A is represented below.



Contrarily to implementation A, in implementation B, MNOs only share aggregated information with third parties. In this implementation, every MNO will provide a modified version of the mobility matrices developed by [quantifying] to the third party. Using three months of data, every MNO will assign every of its users to one region. This region will be the user's home location. The MNO will provide the third party with a region-region matrix containing how much time users whose home is in region i have been spending in region j.

For example, the row corresponding to region i will look like the following matrix:

…	i-2	i-1	i	i+1	i+2	…
…	1%	2%	87%	0.5%	2%	…

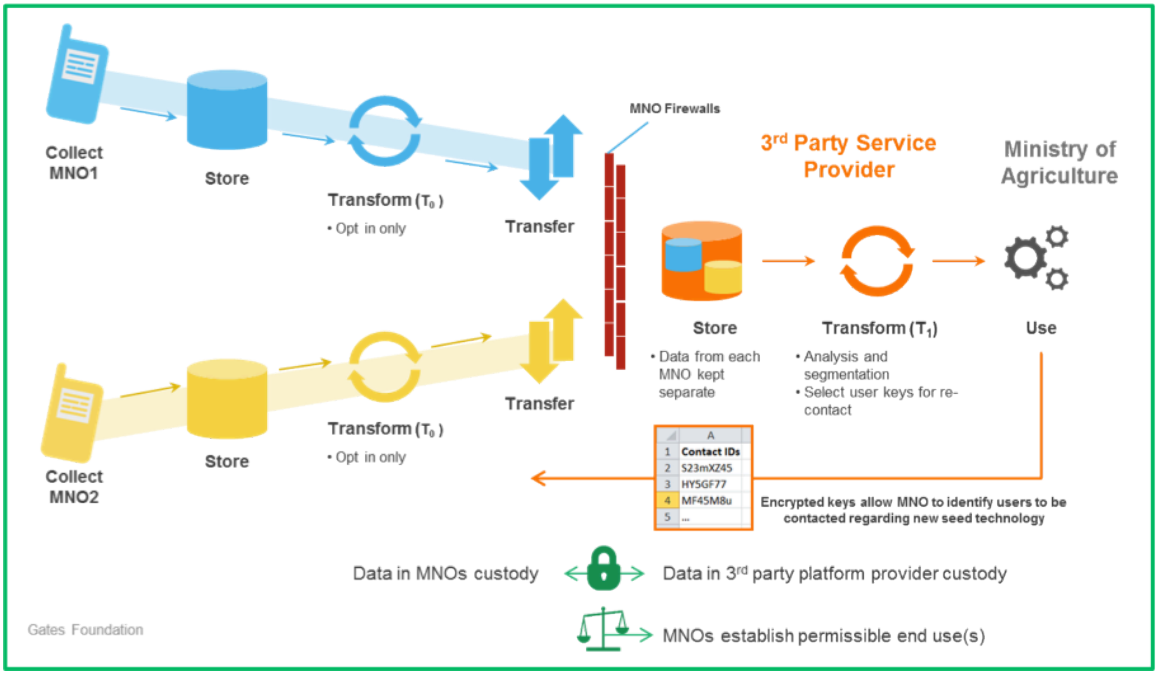
This reads that all the users whose home location is in region i, have been spending 87% of their time (e.g. hourly or nights) in region i, 2% in region i-1, 1% in region i-2 over the course of three months.

Each MNO will also provide the third party with the number of its subscribers who have been assigned to each region.

4.3.2 Scenario 2

Here we will also consider a third party platform provider, although the architecture is fairly similar if there is only the MNO involved. The issue is only that the end users would have to take it upon themselves to link to multiple MNOs if they wanted to be able to target clients of each.

Here the analytic transformation of the data conducted by the service provider would select a set of unique users (not identified by name or other PII, but by encrypted key or other anonymous unique identifier), based on their usage patterns and inferences about their social status or other traits. They would then pass the unique IDs to the MNO, who would be able to match them to the corresponding phone numbers for re-contact with an SMS or automated voice message encouraging program participation.



- Case 1 – Third parties may analyze anonymous data to select individuals, but the mobile operator is the only one in touch with targets and they are not identified to third parties. Third parties may pass back an encrypted key or other identifier to trigger sending a message.
- Case 2 - A third-party is put directly in touch with the targets, or can identify them itself.

4.4 Regulatory Environment

Review of online sources on data privacy laws in Africa indicates a landscape that is evolving along two lines. Francophone countries in West Africa and North Africa that reflect the French civil code system have tended to adopt privacy frameworks modeled on the 1995 European Privacy Directive, supervised by data protection authorities. English-speaking countries with common law systems have less defined privacy laws.

Thus, data protection authorities in a number of French-speaking countries around the world have united in an association under the leadership of the French CNIL, and at least

Benin, Burkina Faso, Gabon, Ivory Coast, Senegal, Madagascar, Mali, Mauritius, and Morocco have such privacy regimes in place, with new laws expected in Mauritania and Niger. Many countries (e.g., Côte d'Ivoire) in both categories do not have any data protection laws, but do appear to have constitutional provisions for a right to privacy that provides at least some authority for protection.

In the English-speaking countries, the systems are less developed. South Africa recently adopted legislation, the Protection of Personal Information Bill that adopts privacy principles to be enforced by a data protection authority; it takes effect at the end of this year. Both Nigeria and Kenya are considering broader bills that resemble each other.

Based on this framework, we will use the European Privacy Directive (EPD) as a benchmark for civil code countries. We will also look to the [Consumer Privacy Bill of Rights] as a way of exploring its application and developing an alternative framework.[5]

4.5 Data Utility

4.5.1 Scenario 1

Implementation A: In this case, the utility seems close to the situation of having access to the full raw data. Data preprocessing and cleaning is harder to do on coarsened data, as unusual behavior might be hidden by the coarsening (e.g. an unusually high number of phone calls).

Implementation B: In this case, the aggregation that is done at MNO level decreases the utility of the data. Considerations include tracking people across borders, removing dual simmers, and taking specific periods of time into account.

4.6 Privacy

Implementation A: There exists a risk of re-identification even when the data is coarsened. We will look at the number of antenna over several regions to match to the unicity formula on spatial resolution. Similarly, the temporal resolution here would be twelve. This should allow a very rough estimate of the likelihood of re-identification given x points.

Implementation B: When data is aggregated, the risk of re-identification is lower; the edge cases would be very small regions that have been assigned as home regions to very few people. The risk to consider here would be at the group level, e.g. people from one region that only go to another region (of the same ethnic group, for example). A counterpoint would be people who spend too much time in another region. This goes beyond pure privacy as risk of re-identification and many other cases should be considered.

[5] Craig Mundie, in a recent Foreign Affairs article, suggests a new model where governance and regulations should not be focused as much at the point of collection and storage of personal data, but rather on how that personal data is used and retained. The President's Council of Advisers on Science & Technology (of which Craig Mundie is a member) echoed many of the recommendations and thoughts. In their document, Big Data: Seizing Opportunities, Preserving Value, in particular, the belief that regulating use cases and enforcing privacy with stiff contractual obligations and deterrents may be needed to extract value while maintaining data security and privacy.

4.7 Critical Issues

- Business case for mobile carriers. Mobile carriers are not in the business of conducting social science or public health research. NGOs will need to develop a business plan that makes data-sharing work for the carriers interesting and worthwhile from their perspective. Support of governments (e.g., health ministries and communications regulators) will be pivotal.
- Scenario 1 presents technical issues of de-identification. The spatial and temporal coarsening of call detail records (CDRs) substantially mitigates privacy risks and, if strong enough, can sidestep the application of the EU Privacy Directive. However, it can also limit the reliability and utility of the data.
- In Scenario 2, de-identification, at least for significant applications, is not an option, because interventions will targeted to specific individuals. This scenario will require engagement of governments to enable the data use and identification; without affirmative support by relevant health and data protection authorities, this scenario may be impossible. The implication of governments will also require careful development of mechanisms to avoid misuse and unwanted identification.
- Further development of specific practices and technical methods to manage privacy protection in accordance with various principles of the EU Privacy Directive and the Consumer Privacy Bill of Rights (e.g. data retention, accountability)

4.8 Promising Paths Forward

Across both of these scenarios there are promising paths forward in terms of employing different technical architectures and practices to meet data protection needs, while still extracting value from the data.

4.8.1 Scenario 1

In this case, there are already private sector companies that grab mobility data from mobile operators and sell it without user permission (i.e., based on ostensibly achieving anonymity).

Airsage is an example in the U.S. that demonstrates a number of innovative approaches to sharing anonymous mobility data. They improve the quality of the position signal over what a CDR would be able to provide through triangulation, which they achieve by upgrading the base station software of the MNO. They then install software within the MNO firewall that anonymizes the data by stripping it down to just mobility patterns and aggregates the output to a minimum of seven mobile traces per observation. Hence, if two people moved from A to B in a given time period, they would report that "less than seven people moved." The fact that they do their anonymization within the firewall removes the need to share raw data.

A company called Grandata in Mexico uses a form of differential privacy algorithm to add some random noise and limit the fidelity of queries on their mobility data that they sell to retail marketers.

Other techniques to explore further would include emerging differential privacy approaches, as well as synthetic dataset generation via modeling methodologies (e.g. DP-WHERE).

4.8.2 Scenario 2

Because decisions are being made about actions involving individuals or small groups in this scenario, and because individual level data (rather than aggregate) are being used, the fact that data is anonymized by being stripped of PII does not fully ameliorate privacy concerns.

Some approaches to investigate here are:

- ID key encryption schemes and anonymization approaches that go as far as possible to protect individual identity.
- Some form of regulatory exception (e.g. specific legal authorization or public policy exception) might also be in order, since even fully anonymized data would still refer to individuals.
- Development of ethical principles to make sure that decisions being made about individuals are fair and do not explicitly disadvantage anyone.
- This requires careful thinking about the user experience – SMS or calls that are clearly targeting the person might feel "creepy" and care should be take not to make data subjects feel uncomfortable or targeted in any way
- The development of trust frameworks to manage the data and verify the legitimacy of its uses

4.9 References

4.9.1 Overview of African Privacy Regulation

[D4D] http://arxiv.org/abs/1407.4885

[UN] http://www.unglobalpulse.org/Mobile_Phone_Network_Data-for-Dev

[unique] http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

[quantifying] http://www.sciencemag.org/content/338/6104/267.abstract

[bigdatadriven] http://web.media.mit.edu/~yva/papers/sundsoy2014big.pdf

https://docs.google.com/document/d/1tSJSADw41Ym-HAjQB9hCgc1s7KnTAB7P_jhEjteP-As/edit?usp=sharing

4.9.2 Scenario development document

https://docs.google.com/document/d/1YG6w5althPpw8KOEIGti_sr9LrOtZBnw8W9eULTkP20/edit#heading=h.gjdgxs

5 Additional Use Cases

Summarized by Karen Sollins (MIT)

In addition to the three scenarios developed above, four other groups provided briefer reports. They are summarized here, in order to further broaden our understanding of the breadth of the problem domain of consideration of privacy in the world of Big Data. These additional topics are: (1) Privacy in Aggregated Diverse Data Sets, (2) Creation, Management, Application, and Auditing of Consent on Personal Data, (3) Consumer Privacy/Retail Marketing, and (4) Genomics and Health.

5.1 Privacy in Aggregated Diverse Data Sets

Team: Evelyne Viegas (Microsoft), Micah Altman (MIT), Yves-Alexandre de Montjoye (MIT), Elizabeth Bruce (MIT)

Overview

Microsoft is working with the research community on developing an open source platform for hosting data sets and code for the machine learning research community. CodaLab is a Machine Learning Service that allows researchers to share and browse code, data, and create and share experiments and workflows. CodaLab helps nurture an environment of scientific rigor and open up new avenues for collaboration between researchers.

The characteristics of data that are submitted might vary widely. Such data includes well-known, previously published data, such as that from official statistics and community-managed data obtained from third parties, data collected by the authors of the submission generally for their research, and derivative datasets prepared specifically for a publication which may integrate, correct, annotate, and recode data from multiple sources.

The emerging challenges in this area are related to the variety of data and the limited resources that are available for vetting it. Owners of community repositories are particularly concerned with developing policies that 1) are strong enough to strengthen replicability, 2) that can be applied without intense case-specific scrutiny, and 3) recognize common disclosure of threats, while still permitting posting and access.

Stakeholders

Data collector - wide range -- any party that collects original data, no direct interaction with service or main scenarios, may have set terms under which data was originally collected

Service host - provides CodaLab service and hosts storage, may impose restrictions on use

Data subjects - wide range -- no direct interaction with service or main scenarios

Data curator - curators create "competitions" on the site, provide data to the service, set terms of use that are presented to competitors, (optionally) vet competitors

Data analyst - entrants in a particular competition, typically researchers who aim to develop or tune algorithms or models to optimize some quantitative competition criteria, such as % correctly predicted, Mean-squared error (MSE)

Data users - synonymous with data analysts

Questions and challenges

Key goals:

- Share research showing advancement in field (not just incremental advances)
- Find experts who can work on a (societal) problem

Key risks:

- Re-identification attacks
- Inadvertent disclosure of personal information

Identified challenges:

- What is the data life cycle?
- How does a service owner manage privacy-related risks resulting from running a service that accepting data from curators?
- Low-effort methods -- must apply to many different data sets of heterogeneous types without expert analysis of each database
- Reuse across challenges: most competitions do not support reuse across challenges, or long-term access. In contrast, a goal of CodaLab is to contribute to a long-term evidence base for research in this area.
- Automatic or guided identification of PII/data currently focuses on medical/health data cases and may not be appropriate to the range of data being considered in this use case.
- How do we measure tradeoffs between utility vs. privacy in this use case?
- Are there automated techniques for identifying potential PII in data sets being submitted by researchers?

5.2 Creation, Management, Application and Auditing of Consent on Personal Data

Team: Simon Thompson (BT), Karen Sollins (MIT), Arnie Rosenthal (Mitre)

Overview

Personal data has many stakeholders. This scenario focuses on the ability of the subject, as an important stakeholder, to influence how their data is *treated:* collected, shared, used, and protected, and the ability of the controllers of personal data to abide by these preferences. Patients and other stakeholders must have incentives to share (and minimize disincentives), and to trust others to behave as they say they will do. Otherwise, patients may withhold data from clinicians and record holders will resist forwarding data to others, harming patients' health, increasing costs, and slowing operational improvements and research progress.

Personal data is of many kinds, often requiring different policies. These distinctions in kind are multi-dimensional, and no single distinction dominates. We note that audit metadata and the subject's own consent specifications are themselves personal data. They do not require fundamentally different treatment, but may have some specific policies attached.

This scenario is relevant to many important verticals, including several each in Healthcare, Education, and Commerce, but what is central to this scenario is the interplay among stakeholders' wishes. These depend on the kind of information involved. In particular, the subject may have different rights with regard to different kinds of data, and especially in terms of medical content.

A critical aspect of this arena is that stakeholders, especially the subjects, deserve *appropriate* controls, but can rarely handle the technical complexity of specifying them. They need a way to customize behavior to be approximately correct. The regulatory framework may need to allow for situations where the user did not specify or understand all behavioral details (just as it allows signoff on legalese that few citizens understand).

Stakeholders:

The key stakeholders consider in this review are:

Data Subjects: those described by the data

Record holders: the collectors and repository

Recipients: those who may receive the data, including, for example with medical records, caregivers, payers, researchers, marketers, or legal authorities, who then may become record holders.

Questions, challenges, and observations:

Key Goals:

- Provide subjects with appropriate (to themselves) understanding and control (user preferences) over privacy policies of information about themselves.
- Balance the interplay between interests and responsibilities of different stakeholders, for example the subject, regulators, caregivers, insurance companies, etc.
- Tagging or other labeling and governance of data in order to enable application of policies.
- Certifying and maintaining the quality of the data

Key Challenges:

- **Preference data is itself metadata about the subject:** Consumer preference data must be tagged by what content the preference itself reveals - a patient preference about releasing abortion data should itself tagged as abortion-related, and cannot be shared with all record holders. It is an open question how best to combine confidentiality and usability for such data.
- **Standards for composition when global standards are impossible:** Global standards, globally complied across all industries, are unlikely – especially as one adds more and more details. (A few basic practices might be standardized and complied with, but not the diversity in a modern economy). How should stakeholders express policies that are robust, even when some information is absent?
- **The diversity of enforcement mechanisms will complicate implementation**: Techniques for a major corporation may be inappropriate for a small business and techniques suitable for managing large documents may be inappropriate for

millions of values in a database. For example, omitting a document differs from redacting a database value (whose absence may be noted).

- **Trust:** To provide an effective privacy management mechanism, the privacy metadata of personal information must be trusted, and used by trusted components, i.e., one needs an effective trust network that assures that everyone will behave appropriately.

5.3 Consumer Privacy/Retail Marketing

Team: John Ellenberger (SAP). Ilaria Liccardi (MIT), Dazza Greenwood (MIT)

Overview:

This group considered a specific example in marketing, a customer loyalty program in a brick and mortar retailer. They envisioned a system with three elements: (1) the customer's smart phone, (2) a cloud-based intermediary service, and (3) the retailer's backend. The intermediary service provides the service for communication with the smart phone, both collecting data and pushing offers. The retailer's backend collects, manages and utilizes the customer data and as part of that provides the support for any privacy policies and meets any legal requirements for privacy.

As an extension to this, the group also considered a case where third-party data may become available to the backend service. The group considered the problem of mapping between the "identified" data collected by the retailer and the potentially anonymized data from a third-party marketing firm.

Stakeholders:

- Subject
- Cloud service provider
- Retailer running the backend data collection, management, and analysis services
- Possible third-party marketing data source

Key goals:

- Improve the customer experience in the store
- Increase the retailer's market share
- To the extent there are regulatory requirements on privacy policy enforcement, comply with the law

Key challenges:

- Fusion of identified data, legitimately collected by the retailer with third-party "marketing data". Simply fusing these correctly is extremely difficult.
- To the extent that merging data may create "new data" about the subject, this is subject to regulations, especially in Europe and it will require permissions from the subject.
- Inference of other facts about a subject from the base level information. For example, it is well understood that patterns of "likes" may be a good predictor of preferences not directly exposed and therefore subject to privacy policies. The demonstrated example is prediction of sexual preferences.
- More broadly this group did not consider the ethics of these approaches.

5.4 Genomics and Health

Team: James Williams (Google/University of Toronto), Michael Power (Osgoode Hall Law School)

Overview:

This scenario focuses on sharing health information (including genetic information) for both health-related research and personalized medicine. The scenario involves numerous health care providers (e.g., hospitals) and research groups (e.g., universities) collaborating to exchange information for a variety of purposes, including the provision of care. As a result, it is inherently complex; not only are there numerous organizations involved, but each of these may be subject to different legal requirements based on the jurisdiction (e.g., country, state, province) in which they operate.

While advances in genomic research methods have major ramifications for the biological sciences in general, they are particularly interesting from the standpoint of *health-related research*. In fact, some researchers have argued that the analysis of large genomic databases (i.e., containing millions of samples, as opposed to thousands) may be the key to unlocking new discoveries related to human health. To name but two advantages: 1) larger data sets empower researchers by supporting a wider range of queries and observations, and 2) the use of modern, distributed computing infrastructure supports interactive modes of research that offer major advantages over traditional approaches.

The situation becomes even more pressing when one realizes that many research problems can only be answered by combining genotype and phenotype data. In practice, this means the merging of genomic repositories with *electronic medical records* (EMRs). Indeed, the emerging field of personalized medicine is based on the ability to correlate information between these two domains. Given the multitudes of health-related issues facing human populations, and the promise of genomic research and personalized medicine to address a significant number of them, it is important to develop tools and methods for fostering the sharing of genetic and phenotypic information for research purposes.

Of course, *privacy* is one of the most commonly cited concerns that arise when individuals are surveyed about their attitudes towards sharing health information. It is vital that such data sharing be accomplished in a manner that minimizes risks to privacy. As part of respecting privacy, individuals must be provided with the ability to control the use of their information, including withdrawing consent.

While informational privacy concerns are explicitly addressed in data protection law, fair information practices, and data sharing agreements, it is an open question as to whether we can design better mechanisms to give effect to these norms.

Stakeholders:

- Patients, subjects of the data
- Clinicians including both physicians and allied health professionals
- Researchers
- Health care service providers
- Institutional Review Boards (IRBs) or Research Ethics Boards (REBs)
- Regulators

Key Goals:

- Delivery of timely and effective health care (patients, clinicians)
- Participate in research (patients, possibly clinicians, researchers
- Act in accordance with "fiduciary" responsibility (clinicians)
- Obtain and utilize large genomic data sets (researchers)
- Obtain and utilize large clinical (i.e. phenotype data) datasets (researchers)
- Integrate across these two types of datasets (researchers)
- Maximize efficiency of healthcare delivery (health care service providers)
- Utilize (and profit from) intellectual property inherent in patient records (health care service providers)
- Maintain security of records systems (health care service providers)
- Minimize privacy risks (regulators)
- Provide recourse for privacy violations (regulators)

Key Challenges:

- At present, integration is almost impossible. Most dataset access is restricted to people within the organization collecting the data.
- Integration across different regulatory authorities is poorly understood.
- The trade-offs between privacy and utility in the context of technical privacy preservation mechanisms are particularly acute in the case of genomic research.
- There is also a tension between the ability of patients (data subjects) to control the use of their information, and the ability of researchers to accumulate stable data sets for research purposes. For instance, dynamic consent mechanisms give patients control of data at the expense of researchers, whose activities may be interdicted by requests to remove data from their corpus.
- Enabling inter-jurisdictional transfer of data may require the harmonization of regulatory regimes, as well as the adoption of common standards.
- The current transaction costs for data sharing agreements are onerous for many organizations, creating a landscape of 'silos' of health information that have great utility, but which cannot be accessed.
- Existing approaches to sharing health data between organizations rely heavily upon bi-lateral data sharing agreements. This approach scales poorly when there are multiple organizations that wish to jointly share data.

6 Conclusions

Karen Sollins (MIT)

We generalize three sets of conclusions from the review of the scenarios described above in Sections 2 through 5. The first is a set of overarching challenges derived from the systemic approaches taken across these Big Data scenarios in consideration of privacy. The second is a common although not universal set of types of stakeholders in handling both the Big Data itself and in support of the application of privacy policies. Finally, we observe a number of key open questions, raised by the set of scenarios.

We observe five key challenges from the scenarios:

- *Scale*: Not only are we observing increasing sizes of datasets, but also those increases in size will lead to increases in size of the accompanying meta-data that is critical to the support of privacy. Without significant improvements in efficiency, the growth in both data and meta-data will lead to untenable processing times, but this must be achieved without cost to privacy.
- *Diversity*: With increasing dataset sizes will also come an increase in interests and types of responsibilities. This increase is likely to lead to increased probability of non-aligned interests. This diversity of objectives and interest will lead to at least a divergence of privacy policies and more likely to increased incompatibility of privacy policies. Capabilities for both observing and handling such differences will become increasingly important.
- *Integration*: In addition to the points above of scale and diversity, services increasingly support the integration of previously independent datasets. At a minimum this can lead to surprising or unintended inferences across these newly integrated datasets, resulting in previously unknown facts about subjects. Thus a new challenge arises from this integration in terms of privacy policies for these newly discovered facts or data.
- *Impact on secondary participants*: Although data may itself have a primary subject, increasingly there will also be secondary participants or subjects, such as friends, parents, guardians, or by-standers, also reflected in the data. Providing privacy through privacy policies for these secondary participants may be even more challenging than for the primary subjects of data.
- *Need for emergent privacy policies for emergent data*: Integration may lead to emergent, or previously unobservable data about subjects. This newly observable data will also require privacy policies, and it is not clear that those new policies will simply be a derivative of the policies applicable to the underlying original data. It is likely that new, emergent privacy policies will be needed, and the challenge is how those new policies will be created, by whom and under what conditions.

The second set of key observations we derive from these scenarios is a list of types of stakeholders, who play a role in setting, enforcing and mitigating the failure of application of privacy policies. We begin with the subjects of the data itself. In some cases, but not all, they play a role in determining applicable privacy policies. Additionally, a decision-maker, who decides what data to collect and how to handle it may play a significant or central role in setting privacy policies. From there we move to the "handlers" of the data. That data will be collected by some party, and may be separately curated for completeness, accuracy, and so forth by a curator. The data may then be stored, managed, and made available by a data platform provider. It will then be used by a data analyst. All of these last four have access to

the data in one form or another. We have then also identified two addition types of stakeholders, whose roles focus on enforcement of privacy policies and recording or auditing of usage of the data. These two final roles are distinct from each. It is possible to have auditing without enforcement, for either legal or mitigation reasons, if a policy is violated. Enforcement benefits significantly from auditing, but is not dependent on it.

Finally, we recognize that there are many open questions. We highlight four here:

- *Novelty*: Although we identified a number of challenges above, there remains a question of whether Big Data leads to new and unique challenges in the provision of privacy, or whether these challenges are only more obvious in the Big Data arena.
- *Tradeoff*: Each of the scenarios presents a significant benefit. These may be economic, social, medical, and so forth. In addition, each presents risks to privacy, both inherently and perhaps because the situation is still new and not well understood. We must ask how to evaluate the tradeoffs between benefits and risks, specifically to privacy. At this point, we do not even have a metric or spectrum along which to consider this tradeoff, and it is not clear that a single one exists.
- *Harm*: The risk to privacy mentioned above is neither binary nor necessarily stable. This leads to a question of whether and how to evaluate the harm that may result from different choices in the tradeoff space between benefits and risks.
- *Trust*: Trust reflects a willingness among stakeholders to accept vulnerabilities. Thus, we must ask how it is that stakeholders determine their level of trust or mistrust in other stakeholders, with respect to the applicability of privacy policies. This includes both the stakeholders' models of trust, how those relate to people's perceptions of each other, as well as what mechanisms and technologies can provide in support of those levels of trust. Furthermore, one must ask how such trust evolves with time and how that might be supported technically.

It is important to recognize that our observations here are limited. They are based on this limited set of scenarios, and even in that context, may be incomplete. They are presented to give the reader a clearer sense of the sorts of challenges and questions that arise from the intersection of Big Data and privacy.

A. Appendix: Privacy Scenario Template

Team: Simon Thompson (BT) & Dazza Greenwood (MIT Media Lab)

Elements of Big Data scenario

- **People/Stakeholders**? (i.e., Who are the parties, their respective roles and relationships? Who is data owner (data controller)? Who is using the data and what is the intended purpose? Who are the data subjects? Who is doing the data analytics?)
- **Interactions?** (i.e., What transactions or other exchanges between Actors?) (What is the power dynamic?)
- **Data**
    - What kind of personal data?*
    - What type of big data models, analytics, or other outputs result from this scenario?
    - How is the data used?
    - What's the Data Lifecycle?
- **Systems**? (i.e., What business, legal, technical, or social systems matter most?)
    - **Business Systems** (Ethics committees, sign-off by authorized officers, record keeping, audit)
    - **Legal Systems** (Contracts, Employee rules/procedures, certification/accreditations, compliance reviews, insurance/bonding requirements, industry standard policy/guidelines, etc.)
    - **Technical Systems** (System permissions and security, alarms & automated detection of PAI, automatic anonymization of data, cryptography, etc.)
    - **Social Systems** (What social systems and context exists?)

Analysis of scenario

- **Goals** (i.e., What are the incentives and the benefits driving the Actors? Who benefits? What are financial incentives?)
- **Rules**: (i.e., What are the relevant laws and regulations, other enforceable rules)
    - Are there existing statutes, contractual agreements or other commitments associated with the date.
        i. Rules about retention,
        ii. Liability for breach?
        iii. Accuracy?
        iv. Others...
    - If there are not statutory or other binding rules, how would the principles from the Consumer Privacy Bill of Rights guide the development of rules?
        i. INDIVIDUAL CONTROL: Consumers have a right to exercise control

over what personal data companies collect from them and how they use it.

ii. TRANSPARENCY: Consumers have a right to easily understandable and accessible information about privacy and security practices.

iii. RESPECT FOR CONTEXT: Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.

iv. SECURITY: Consumers have a right to secure and responsible handling of personal data.

v. ACCESS AND ACCURACY: Consumers have a right to access and correct personal data in usable formats, in a manner that is appropriate to the sensitivity of the data and the risk of adverse consequences to consumers if the data is inaccurate.

vi. FOCUSED COLLECTION: Consumers have a right to reasonable limits on the personal data that companies collect and retain.

vii. ACCOUNTABILITY: Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.

- **Risks**: What are the potential harms? What are the risks of those harms occurring? To whom? If the risk is an externality, how might it be mitigated?

Assessment of scenario

- **Existing or related best practices for context of this scenario**
  - What business, legal, and/or technical best practices?
- **Gap**
  - Issues Not Addressed by Existing Practices and Solutions
    - **Business Systems**
    - **Legal Systems**
    - **Technical Systems**
    - **Social Systems**
  - Short Fall Between Current and Needed Practices and Solutions
- **Key outcomes for each scenario**
  - Promising best practices
  - Gaps that need to be filled with new tech solutions or policy approaches

- Personal Data is defined broadly, as follows, from the Consumer Privacy Bill of Rights. “This term refers to any data, including aggregations of data, which is linkable to a specific individual. Personal data may include data that is linked to a specific computer or other device. For example, an identifier on a smartphone or family computer that is used to build a usage profile is personal data. This definition provides the flexibility that is necessary to capture the many kinds of data about consumers that commercial entities collect, use, and disclose.

B. Appendix: Stakeholders

Elizabeth Bruce (MIT), Karen Sollins (MIT)

Data Stakeholders	**Decription/Examples**
"data collector"	Party that collects the "raw" or original data from the data subjects
"data subject(s)"	A person (e.g. a patient, student, customer…) or group of people (or entity) that data is being collected from; this is the group of data providers or participants. Subjects may be contributing data with informed consent (e.g. by opting-in to research study); or data may be collected in-directly or in aggregate. Data may be generated by • an individual/consumer (e.g. taking an online class, a customer at a bank) • the interactions of a group of individuals (e.g. peer to peer interactions; social network graphs) • combining/aggregating data over a group/population of subjects.
"data curator" (also: controller, provider or caretaker)	Party that stores and manages the data and is responsible for granting/controlling access to the data; data curator is often the stakeholder that requires others to formally submit to a policy (or Data Use Agreement) in order to access the data. There may be more than one data curator: • original data curator • third-party data curators
"data analyst" (also: data scientist)	Party doing the analytics on the data; may use many different types of tools, software etc for analysis, exploration and visualization ("relying party")
"decision maker"	The stakeholder(s) that benefits from the data; a decision maker that ultimately derives new insights and value from the data analysis; this stakeholder will ultimately make decisions based on the data and may or may not take action for some purpose. This purpose or *use* of the data

	may be for: personal benefit; for-profit or commercial use; or societal benefit e.g. NGOs/government). Data beneficiary may be: • an individual • a group of individuals • an institution or organization (private; commercial; government; non-profit) • a content provider • a service provider
"data platform provider"	The party that builds the system(s) for data collection and provides a service. Platform provider and data collector may or may not be the same entity/organization. In the case that they are different, the Platform provider may have its own Data Use Policy separate from the data collector.
"data regulator(s)"	An arbiter that sets policies; the governing regulatory body that develops policies that controls data collection, sharing and use among stakeholders – could be at the local, state, federal, international level (e.g. HIPPA, FERPA etc)
"data auditor"	The enforcing body responsible for ensuring the policies and regulations are enforced. May require audit logging, documentation – to ensure policies are enforced, and data is managed as required

C. Appendix: Stakeholder Data from MOOCs and Online Learning Environments (OLEs)

Elizabeth Bruce (MIT)

Data Stakeholder	Example
Type of Data	All click stream data capturing interactions between student and content, including when watch video/lessons, quiz answers, text from discussion forums, etc. Use of videos and other e-resources, such as digitized reference material, wikis, and forums. Assessment behavior: attempts, correctness, use of immediate feedback. May include PII (name, email, address) depending on what information required when register for course. Self-reported background, pre and post-test surveys.
Data Subjects	Students who take the online course, complete assignments and receive credit Students have zero or little access to their data beyond official grade/records created for their education purposes
Data Platform Provider	Cousera, EdX, Udacity, Stanford U, etc.
Content Provider	Individual Content Providers include faculty, teachers, staff who provide the teaching content and material (videos, lessons, quizzes, etc), support discussions, interact with students (the data subjects) directly, and responsible for grading/credit Institutional Content Providers include institutions and organizations that are behind the teaching content (i.e. MIT, Harvard, or an individual private enterprise)
Data Collector	Data Platform Providers and Institutional Content Providers
Data Curator	Data Platform Providers and Institutional Content Providers
Data Scientist	Analysts include researchers, their students (if the researchers are academics), and education technologists. Teaching staff, platform providers, and

	institutional content providers may also act as analysts.
Decision Maker	Typically the Data Platform Providers and Institutional Content Providers, sometimes the Individual Content Providers (i.e. the teachers)
Data Auditor (and Compliance)	Government
Data Regulator	Government – FERPA policies